Metodologia da Pesquisa Científica 2026/2

E08: *Comparação de métodos*

Alexandre Valadão Delazeri <alexandre.delazeri@ufpr.br>
David Menotti - web.inf.ufpr.br/menotti
Departamento de Informática
Universidade Federal do Paraná (UFPR)

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Sumário

- Motivação
- O experimento como amostra
- Comparação de resultados
- O intervalo de confiança
- O tamanho de efeito
- Comparações com mais de dois resultados
- Em suma

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Uma simples comparação

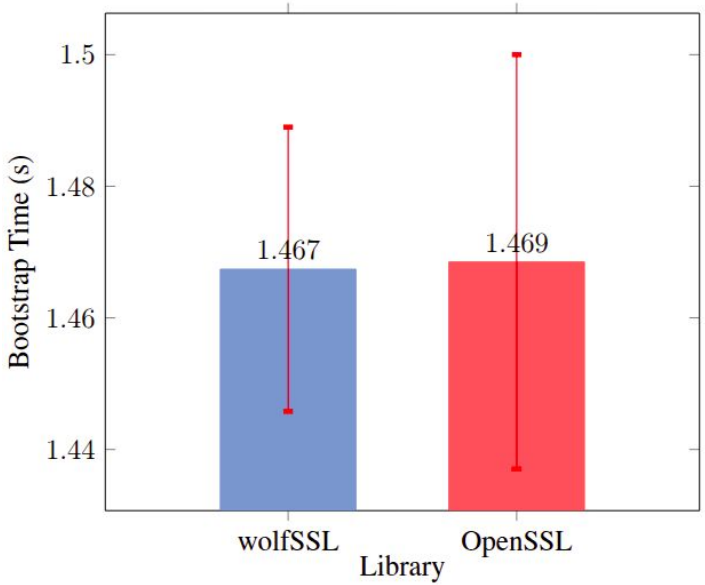


Figura: Tempo médio de execução de dois algoritmos com barra representando o desvio padrão. Figura 7 em [Ehlert et al., 2025].

Figure 7 compares the pledge application using WolfSSL and the one using OpenSSL regarding the total bootstrap time, measured in seconds (s). The results show a slightly faster total time for the WolfSSL implementation. The WolfSSL version executed the bootstrap in an average of 1.467s, versus 1.469s for OpenSSL version (Figure 7a).

Between the pledge applications, the average bootstrap time had a difference of 0.001 seconds (or 1 ms).

Figura: Afirmações na discussão e conclusão de [Ehlert et al., 2025].

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Uma simples comparação

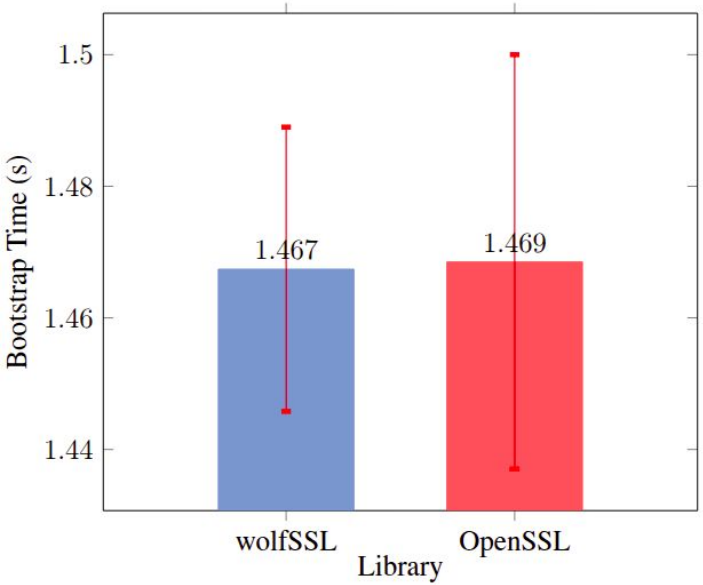


Figura: Tempo médio de execução de dois algoritmos com barra representando o desvio padrão. Figura 7 em [Ehlert et al., 2025].

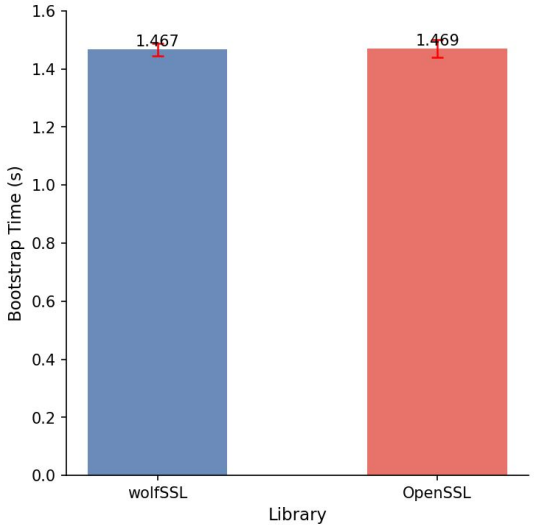


Figura: Mesmo gráfico sem truncagem no eixo, imagem própria.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Objetivo

Análise de resultados

- Fiz um experimento, obtive um resultado. Mas e agora?
- Como comparar resultados diferentes?
- Como comparar meu resultado com resultados de outros pesquisadores?

Dados dois resultados A e B, fundamentalmente existem 3 possibilidades:

- A = B
- A > B
- A < B

Estamos interessados em:

- d = B - A
- A diferença observada bruta

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Sua métrica não é um valor

- Obtivemos um resultado?
- Mas se repetirmos o experimento, o resultado será o mesmo?
- O seu resultado é uma **Variável Aleatória.**

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

A variável aleatória

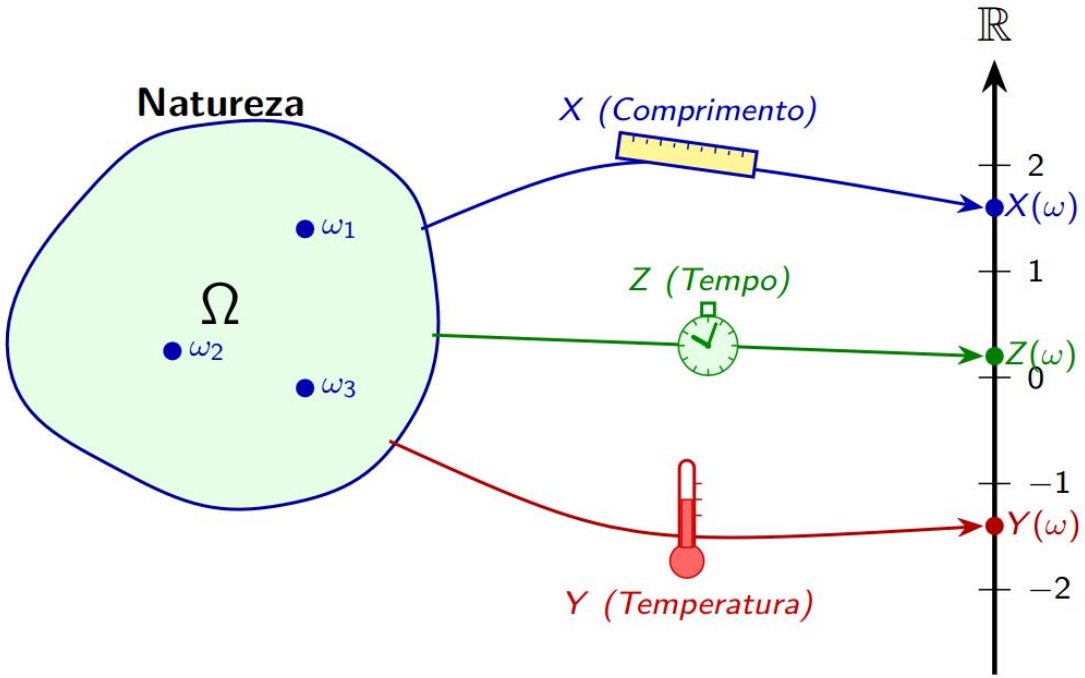
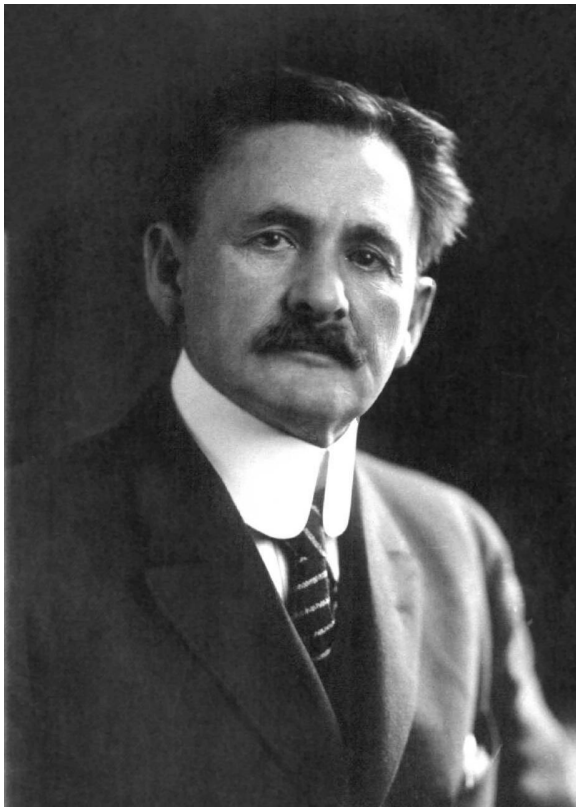
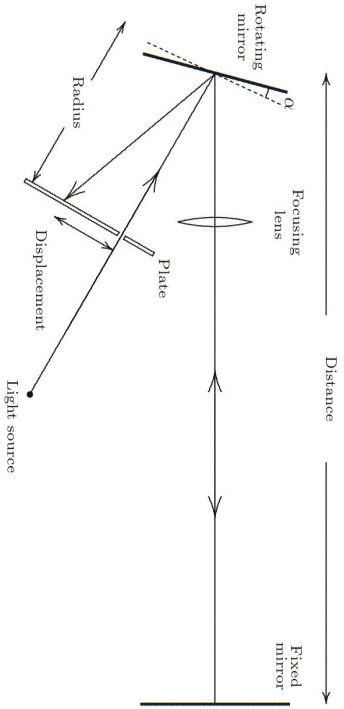


Figura: O conceito de variável aleatória

O conceito de experimento





O conceito de experimento começou a ser formalizado no fim do século 19.

- Cada medida é entendida como realização de uma variável aleatória.
- Cada medição deve ter a mesma distribuição e sob mesmas condições.

Figura Esquerda: Albert A. Michelson
Figura Direita: Esquemática do experimento, fonte: [Dekking et al.].

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

População vs. Amostra

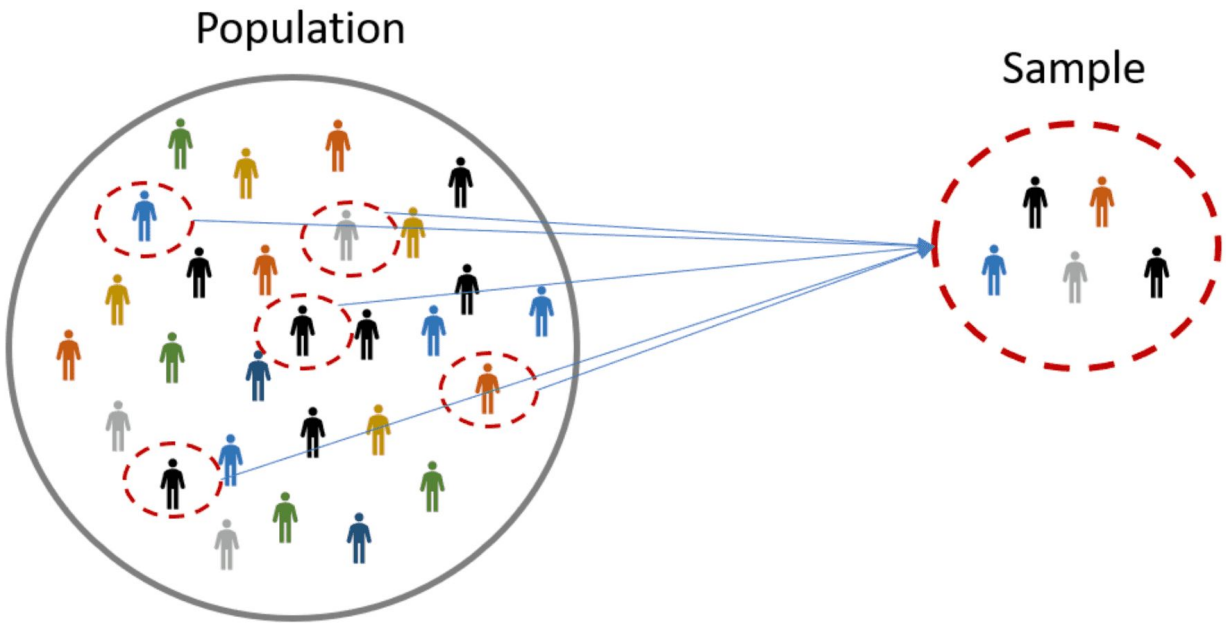


Figura: População e Amostra.
Fonte: Omniconvert

Independentes e Identicamente Distribuídos?

Em geral supomos que as medições são:

- Independentes: Cada processo de medida não interfere nos outros.
- Identicamente distribuídas: As condições são idênticas e as medições todas seguem a mesma distribuição de probabilidade.
- i.i.d.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

O que pode dar errado?

- Independência
  - Ex: dependência entre medições (ex: caching, throttling, carry-over effects).
  - Benchmarks em CPU com thermal throttling.
  - Experimentos sequenciais onde o resultado anterior influencia o próximo.
- Distribuição Idêntica
  - Ex: drift de população.
  - Realizamos uma sequência de medidas sem perceber que o processo muda com o tempo.
  - Pense em um servidor onde medimos requisições sem saber se certa data se refere a um feriado.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

O conceito de erro padrão

- O SE (do inglês *Standard Error*) é uma medida da dispersão da estimativa
- SE = desvio padrão da estatística
- Para média, SE = $\sigma/\sqrt{n}$
- σ = Variabilidade dos valores individuais
- Maior σ -> Maior SE
- Maior n -> Menor SE
- Na prática não temos acesso a σ e utilizamos S, o desvio padrão da amostra, para estimá-lo.

Visualizando

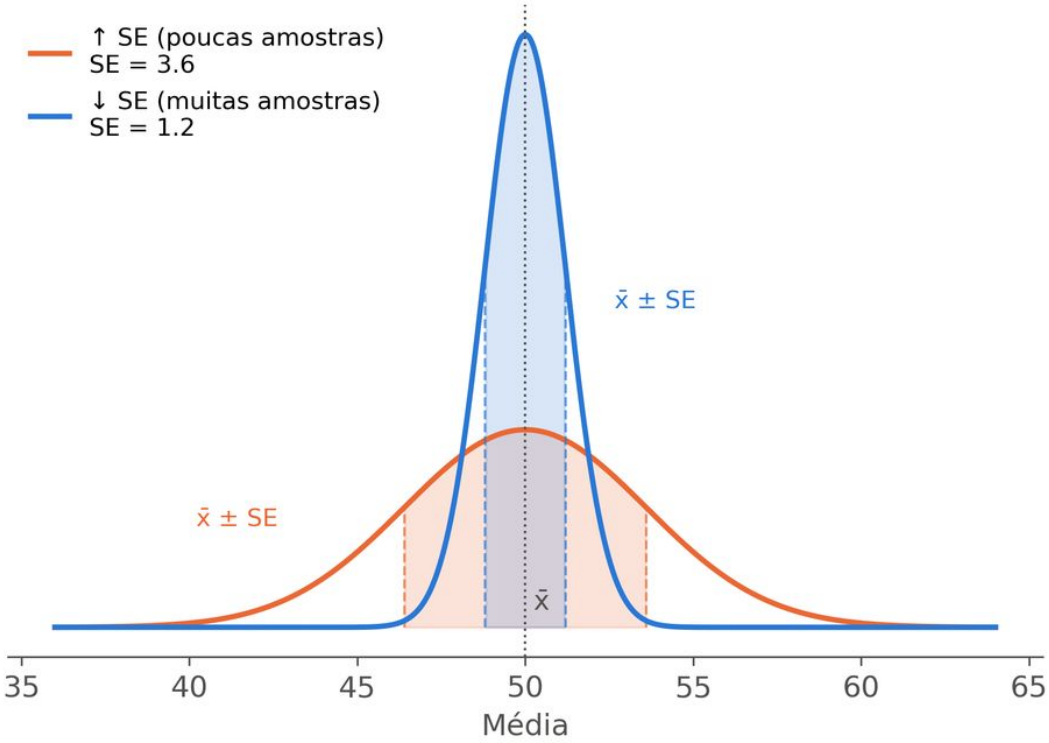


Figura: O SE é uma medida da dispersão do valor estimado ao entorno do valor real.

Ok, mas e agora?

- Nosso experimento é apenas uma amostra, ok.
- Observamos duas médias diferentes, 1.467 e 1.469.
- E o SE?
  - Dando todas as colheres de chá possíveis, 0.00221
- Fundamentalmente existem duas possíveis razões para isso:
  - Existe uma diferença real entre as duas populações.
  - Variação natural da medida devido ao processo amostral.
- Se repetirmos o experimento, obtemos um novo resultado.
- Como distinguir entre esses casos?

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Hipóteses

- A hipótese H_0
  - Assuma, por hipótese, que não há diferença entre as amostras. Queremos saber o quão surpreendentes são os nossos dados.
  - Intuitivamente esse é o pior caso. Se quero mostrar que algo existe, parta do pressuposto que não existe.
- A hipótese alternativa H_A
  - O efeito que acreditamos que exista.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

O teste em si

- De maneira generica:
  - Calcula-se a estatistica de teste (e.g. a média)
  - Compara-se a estatistica com um valor critico determinado pelo procedimento específico
- Para o caso da média:
  - $|d/SE| > t$
- Geralmente, o valor t é selecionado tal que a chance de erro seja 5% a priori
- Mas quão surpreendente é o resultado observado se H_0 for verdadeira?
  - O p-valor quantifica essa surpresa.
  - Quanto menor o p-valor, menos compatíveis são os dados com H_0.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Por que 0.05?

“O valor para o qual P = 0.05, ou 1 em 20, é 1.96 ou quase 2; é conveniente então tomar esse valor como o limite para julgar se um desvio deve ser considerado significativo ou não. Desvios excedendo duas vezes o desvio padrão são então formalmente considerados como significantes. Usando esse critério, seríamos levados a seguir uma indicação falsa apenas uma vez a cada 22 tentativas, mesmo que as estatísticas fossem o único guia disponível.”

– Fisher. R, 1925, Statistical Methods for Research Workers[1]

1. Tradução própria.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

A conveniência de Fisher

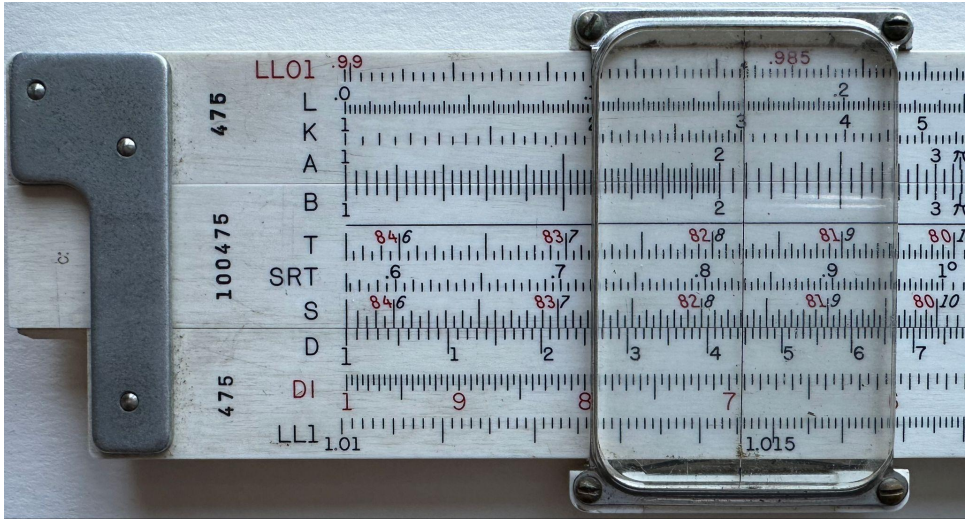




Figura: O estado da arte das ferramentas para fazer contas em 1925.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

O que é o p-valor?

- O p-valor é uma **medida de compatibilidade** entre nossos dados e a hipótese nula
  - Dados pouco compatíveis com a hipótese H_0 fornecem evidência de que ela possivelmente seja falsa.
- Definição
  - Probabilidade de que, dado que a hipótese H_0 seja verdadeira, sejam observados dados tão (ou mais) extremos quanto os nossos.
- Note que o p-valor é sempre referente a uma hipótese nula e um modelo que estamos assumindo (com todas suas suposições).

O que significa ter “evidência estatística”

- A evidência é um gradiente, não uma linha de chegada
  - “Evidência” é o grau de surpresa ao observar nossos dados, assumindo que a hipótese H_0 seja a realidade.
- Em um mundo onde H_0 seja verdadeira:
  - p-valor alto ($p > 0.10$): Os dados não são surpreendentes.
  - p-valor pequeno ($p < 0.05$): Evidências de que os dados são incomuns.
  - p-valor minúsculo ($p < 0.001$): Os dados seriam extremamente bizarros.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Exemplo de cálculo de p-valor

- Suponha que tenho alguma razão para desconfiar que uma dada moeda é viciada.
- Tomo H_0: a moeda é justa, H_A a moeda tem uma chance maior de resultar em caras.
- Meu experimento: Jogo a moeda 30 vezes e conto o número de caras, C.
- Digamos que observei 20 caras, isto é 2/3s dos lançamentos resultaram em caras. $C_{obs} = 20$.
- Qual o p-valor?
  - $p = \mathbb{P}(C \geq C_{obs} \mid H_0) = \mathbb{P}(C \geq 20 \mid H_0) = \sum_{k=20}^{30} \binom{30}{k} (0.5)^{30} \approx 0.0494.$
- Significante?

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Interpretando o exemplo

- Possivelmente.
- Por um lado, 2/3s dos lançamentos serem caras parece inusitado.
- Por outro lado, isso ocorre aproximadamente 1 em cada 20 vezes.
  - (1/0.0494 ≅ 20.24)
- No fim, vivemos neste mundo, onde coisas inusitadas podem acontecer de vez em quando, não no “mundo médio” onde apenas coisas típicas ocorrem.
- Mas o teste pode servir de razão para buscarmos mais evidência.

De volta ao openSSL vs wolfSSL

- Não temos acesso às medidas de tempo de cada execução.
- Porém podemos considerar o melhor caso possível:
  - Assumir que as medições foram ideais
  - Supor que todo o ruído externo afetou os dois algoritmos de forma idêntica
- $$\frac{d}{SE} = \frac{1.469 - 1.467}{\sqrt{\frac{0.0216^2 + 0.0315^2 - 2 \cdot 0.0216 \cdot 0.0315}{20}}} = 0.9035$$
- p ≅ 0.3776
- Ou seja, se ambos fossem iguais observariamos resultado como esse 1 em cada 3 vezes.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Regra Pratica #1

- Guarde os dados tão completos quanto possível
  - Não reporte apenas as medidas de interesse.
  - Guarde os dados brutos e a ordem em que eles foram observados.
  - Essas informações podem ser úteis em análises futuras.
  - Sempre que possível reporte o dataset coletado e não apenas as estatísticas de resumo.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

O que NÃO é o p-valor?

- Probabilidade que H_0 é verdadeira
- Probabilidade de que o resultado seja devido ao acaso
- Proporcional à magnitude do efeito
- Medida de importância do resultado
- Evidência direta em favor da hipótese alternativa
- Resposta definitiva se sim ou se não

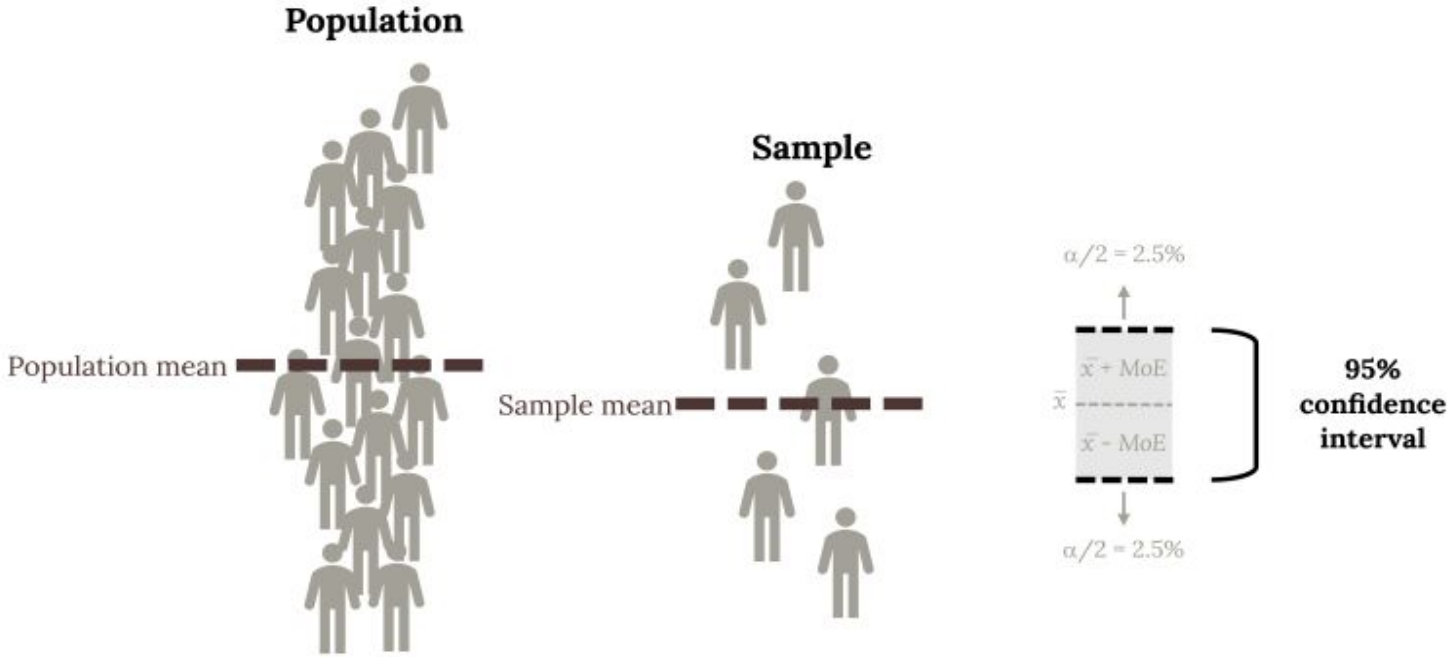
UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Qual a incerteza associada à nossa medida

- Se, ao invés de comparar medidas pontuais, quiséssemos saber quais são os valores compatíveis com o nosso estudo?
- O intervalo de confiança é uma ferramenta que visa ajudar a responder essa pergunta.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Interpretação gráfica



Intervalo de confiança

- O intervalo de confiança fornece uma faixa de valores compatíveis com os dados observados
- A teoria diz que Margem de Erro $\propto$ SE
- Mas como determinar a constante de proporcionalidade?

A “tal” da constante de proporcionalidade

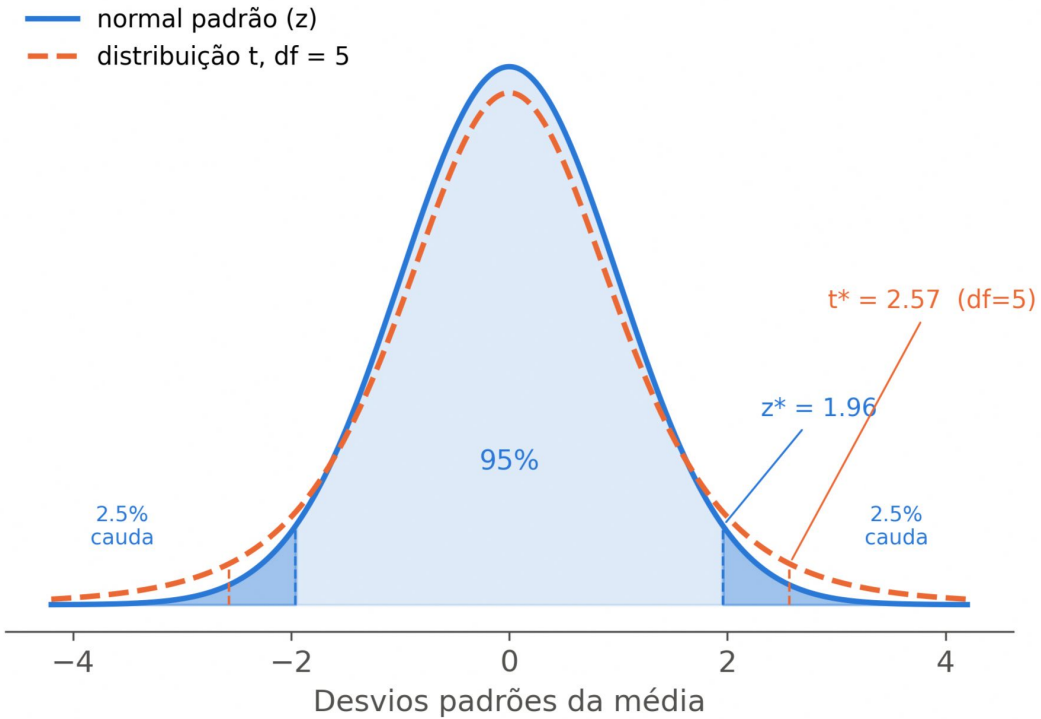


Figura: A constante de proporcionalidade é o valor z que garante 95% da área sob a curva entre −z e +z.

Para a normal este valor é z = 1.96

Exemplo: Calculando um intervalo de confiança de 95%

- Em 20 amostras do wolfSSL: $\bar{x} = 1.467$, $S = 0.0216$
- Para confiança de 95% e 19 graus de liberdade, o t-score é 2.09.
- Dai $\bar{x} \pm t_{\alpha/2,n-1}SE$.
- Substituindo: $1.467 \pm 2.09 \cdot 0.0216/\sqrt{20} = [1.457, 1.477]$
- Para o openSSL, temos $\bar{x} = 1.469$ e $S = 0.0315$
- Dai $1.469 \pm 2.09 \cdot 0.0315/\sqrt{20} = [1.454, 1.484]$

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Um intervalo de confiança para nossos dados

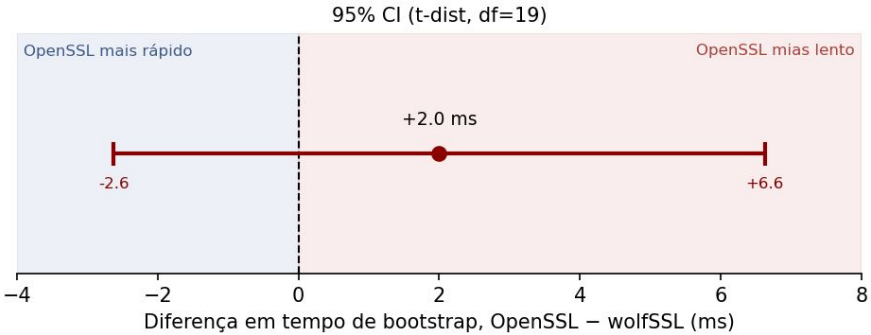


Figura: Intervalo de confiança para os nossos dados

- Valores da média dentro de [1.457, 1.477] e [1.454, 1.484] respectivamente são compatíveis com nossos dados.
- O estudo não é capaz de descartar valores fora dessas faixas.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Mas o que X% de confiança significa?

- Isso significa que o intervalo foi construído por um método que, se repetido muitas vezes, conteria o valor verdadeiro em aproximadamente 95% dos casos.
- Importante: isso não significa que existe 95% de probabilidade de o valor verdadeiro estar nesse intervalo. A média é um valor específico fixo. É a **média da amostra** que depende da amostra.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Interpretação grafica

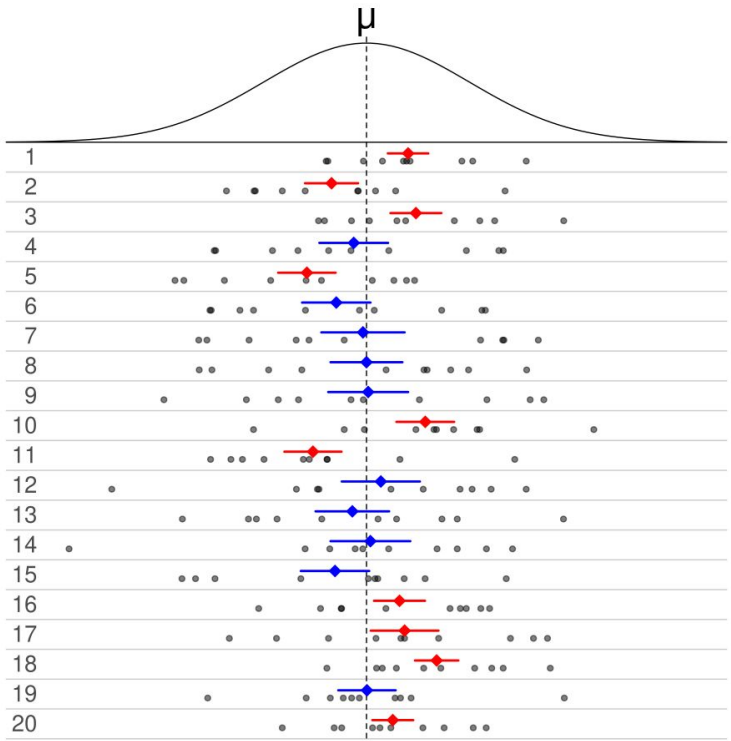


Figura: Simulação com 20 amostras de uma distribuição normal com intervalo de confiança de 50% para a média. Em geral, esperamos que metade dos intervalos contenha a verdadeira média. Na imagem ao lado, 10 intervalos contem a média.

Fonte: Wikipedia.

O intervalo de confiança contendo zero

- Como vimos anteriormente, o intervalo de confiança é o conjunto de valores que nosso estudo não é capaz de descartar.
- Logo, para comparações, um intervalo de confiança da diferença contendo zero sugere que o estudo é incapaz de descartar igualdade.
- Por isso esta é uma heurística comum para avaliação.
- Novamente no melhor caso:
  - $0{,}002 \pm 2{,}093 \cdot 0{,}00221 = [-2{,}6; +6{,}6]$ ms
  - Assim o estudo suporta desde o openssl ser 2.6ms mais rapido até o 6.6ms mais lento
  - Novamente, estatisticamente não podemos concluir nada dos dados reportados

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Regra prática #2:

- Sempre que viável, reporte um intervalo de confiança
  - *Um único valor, sem estimativa de incerteza, não é muito significativo.*

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Tamanho de efeito:

- Relevância prática
  - Mesmo se algo ocorre, é algo importante?
- Pense na comparação de implementações do processo de autenticação:
  - Diferença observada: 2 milisegundos. Isso importa?
  - Talvez
    - Servidor: 100000 autenticações/dia → 3,3 minutos economizados.
    - Dispositivo IoT: 1 autenticação/dia → irrelevante.
  - Vale a pena investigar?

E se tenho mais de dois resultados?

- O problema dos pares
  - Com k metodos temos k(k-1)/2 comparações
  - K = 5→10, K = 10→45, K = 20→190
- Se eu comparar 10 pares em que nenhuma diferença é real, quantos ‘achados’ espero?
- P(pelo menos um falso positivo em m testes) = 1 - 0.95^m

m	1	5	10	20	100
$P(\geq 1$ falso positivo)	5%	23%	40%	64%	99,4%

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Correção de Bonferroni

- Note que 23/5 = 4.6, 40/10 = 4, etc
- Valores proximos a 5
- Dai podemos definir a correção do nivel de significancia α/m
  - Ex: se estamos comparando 5 pares, então só consideramos significativo testes com p-valor menor que 0.05/10 = 0.005
- Essa é a técnica de correção mais famosa.
- Novamente, existem técnicas diferentes para hipóteses diferentes.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Testes globais

- Respondem a pergunta: algum método difere dos demais?
- Não indicam qual das medidas é estatisticamente maior/menor
- Comumente é recomendado realizar um teste global e então testar todos contra todos com correção
- Mais famosos: ANOVA, Kruskal-Wallis

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Recomendações práticas

1. Realize repetições suficientes.
2. Colete os dados com informação de ordem.
3. Reporte intervalos de confiança.
4. Reporte o desvio padrão, não apenas as médias.
5. Verifique suas hipóteses.
    a. Hipóteses são aproximações da realidade, quanto menos próximas, maior o erro.
6. Disponibilize os dados coletados como dataset e não apenas estatísticas de resumo.

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Leitura Complementar

1. Ehlert et al, A comparison between cryptography libraries used in BRSKI protocol for constrained devices
2. M. N. Magalhães e A. C. P. de Lima, Noções de probabilidade e estatística
3. F.M. Dekking et al., A Modern Introduction to Probability and Statistics
4. G. E. Noether, Introduction to Statistics: The Nonparametric Way
5. B. Efron, An Introduction to the Bootstrap
6. The American Statistician, Volume 73, Issue sup1 (2019), Statistical Inference in the 21st Century: A World Beyond $p < 0.05$

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ



That’s all folks

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ